An empirical analysis of alleged misunderstandings of coefficient alpha

Hoekstra, R.; Vugteveen, J.; Warrens, M. J.; Kruyen, P. M.

Published in:
International Journal of Social Research Methodology

DOI:
10.1080/13645579.2018.1547523

Publication date:
2019

International Journal of Social Research Methodology

ISSN: 1364-5579 (Print) 1464-5300 (Online) Journal homepage: https://www.tandfonline.com/loi/tsrm20

An empirical analysis of alleged misunderstandings of coefficient alpha

R. Hoekstra, J. Vugteveen, M. J. Warrens & P. M. Kruyen

To cite this article: R. Hoekstra, J. Vugteveen, M. J. Warrens & P. M. Kruyen (2019) An empirical analysis of alleged misunderstandings of coefficient alpha, International Journal of Social Research Methodology, 22:4, 351-364, DOI: 10.1080/13645579.2018.1547523

To link to this article: https://doi.org/10.1080/13645579.2018.1547523

Published online: 18 Dec 2018.

INTERNATIONAL JOURNAL OF SOCIAL RESEARCH METHODOLOGY
2019, VOL. 22, NO. 4, 351–364
https://doi.org/10.1080/13645579.2018.1547523

Routledge
Taylor & Francis Group

An empirical analysis of alleged misunderstandings of coefficient alpha

R. Hoekstra [a], J. Vugteveen [b], M. J. Warrens [a] and P. M. Kruyen [c]

[a]GION Education/Research, University of Groningen, Groningen, The Netherlands; [b]Heymans Institute for Psychological Research, University of Groningen, Groningen, The Netherlands; [c]Institute for Management Research, Radboud University Nijmegen, Nijmegen, The Netherlands

ABSTRACT

Cronbach's alpha is the most frequently used measure to investigate the reliability of measurement instruments. Despite its frequent use, many warn for misinterpretations of alpha. These claims about regular misunderstandings, however, are not based on empirical data. To understand how common such beliefs are, we conducted a survey study to test researchers' knowledge of and beliefs about alpha. For this survey, we selected authors from recent papers, in which alpha was used. The results provide empirical evidence for the claims that researchers have difficulty interpreting alpha in a proper way. At the same time, we expounded the claims, by showing that whereas some beliefs are fairly typical, others are not so often seen. This non-technical paper, aimed at both statisticians and substantive researchers, is concluded by providing a few suggestions that could be helpful to get us out of the current stalemate regarding the usability of alpha and its alternatives.

ARTICLE HISTORY
Received 2 August 2018
Accepted 8 November 2018

KEYWORDS
Coefficient alpha; survey study; misinterpretation; reliability

Tests and questionnaires are instruments that are commonly used in the behavioral and social sciences to measure knowledge, skills and attitudes of participants or respondents. Once a test or questionnaire has been administered, a test score, for example the sum score, is used to summarize the knowledge, attitude or performance of a respondent. A desirable property of a test score is that for each participant, it would produces the same outcome when the questionnaire was to be repeatedly administered under identical test conditions. In test-theoretical terms, this is referred to as high reliability of the test score. A formal definition of reliability comes from classical test theory: reliability is defined as the ratio of the true score variance to the total score variance (McDonald, 1999; Revelle & Zinbarg, 2009).

In practice, the reliability of a test score must be estimated from the data of a study. The measure that is most frequently used to estimate reliability in behavioral and social science research is coefficient alpha (Field, 2009; Furr & Bacharach, 2008; Warrens, 2015). Alpha is a so-called internal consistency reliability measure, which means that its calculation only requires a single administration of a test (Osburn, 2000). Alpha was introduced by Kuder and Richardson (1937) for dichotomous items. A generalized version can be found in Jackson and Ferguson (1941), Hoyt (1941) and Guttman (1945). Alpha was presented as an alternative for the split-half method for reliability. The split-half method was based on the correlation between one randomly or intentionally selected half of the test with the other half, whereas alpha could be considered the *average* correlation for all possible random splits (Revelle & Zinbarg, 2009; Warrens, 2016). A few years after Guttman's introduction, alpha was popularized by Cronbach (1951). The term alpha came from Cronbach, who expected that alpha would just be the first of a range of similar

CONTACT R. Hoekstra r.hoekstra@rug.nl Grote Rozenstraat 2, Groningen 9712 TG, The Netherlands

measures, which could be given subsequent letters from the Greek alphabet. Since this seminal paper, alpha has been used in thousands of research studies (Cortina, 1993; Sijtsma, 2009).

Despite its frequent use amongst researchers, alpha is not without criticism. Sijtsma (2009) wrote about the 'very limited usefulness of Cronbach's alpha' (p. 107). Green, Lissitz and Mulaik (1977) discussed the limitations of coefficient alpha as an index of test dimensionality. Recently, McNeish (2017) wrote a paper with the provocative title 'Thanks coefficient alpha: We'll take it from here', in which he argued that alpha is 'riddled with problems' (p. 1). Cronbach (1951) himself warned for misuse, by acknowledging that it could only be used for tests that are not very short, and for tests that are not divisible into distinct subsets. Others indicate prevalent misuse of the coefficient as well (e.g. Cho & Kim, 2015; Green, Lissitz & Mulaik, 1977; Sijtsma, 2009; Schmitt, 1996). Cortina (1993, p. 98.) talks about a 'widespread lack of understanding', and Sijtsma states '…presumably no other statistic has been subject to so much misunderstanding and confusion' (p. 107).

Of course, this does not imply that there are no people who consider alpha to be a useful measure. McNeish's paper, for example, was heavily criticized in a paper with the telling title 'Thanks coefficient alpha, we still need you' (Raykov & Marcoulides, 2017). Raykov and Marcoulides claimed that McNeish's claims about alpha's lack of usability were premature, and that alpha should still have a prominent place in our statistics toolbox, because under certain assumptions they consider fairly common in practice, reliability and alpha coincide. To be clear, both sides in this debate seem to agree that simply always using coefficient alpha, as seems to be standard practice for many, is unjustified (Cho, 2016). The main difference seems to be that whereas some (like McNeish) would rather get rid of alpha altogether, others (like Raykov and Marcoulides) think using alpha when some conditions are met is warranted, and sometimes even preferable to the use of its alternatives. Moreover, it seems self-evident that there seems no disagreement on the fact that if alpha is used, it should be interpreted correctly.

There is little empirical evidence for the claim that the incidence of these misconceptions is high. Most likely, but this is speculative, the authors claiming widespread misconceptions have often encountered misuse in their scientific environments (when teaching students, consulting for other researchers or reading papers), but for making strong scientific claims, empirical findings are preferable. Given the potential impact of misunderstandings, and given the fact that it is often claimed that alpha is misunderstood, we think a study on researchers' beliefs surrounding alpha is important, but missing. For that reason, we conducted a survey study on researchers' beliefs regarding coefficient alpha. The results presented in this paper provide, as far as we are aware of, the first empirical evidence for the claims that researchers have difficulty interpreting alpha in a proper way. Although this is independent of the discussion on alpha's usability, those who involved in this debate now have more concrete information on how alpha is typically interpreted.

Alleged misunderstandings

Despite there being little information on researchers' use and interpretation of alpha, quite a lot has been written about specific misinterpretations of or undesirable beliefs about alpha among researchers. In this paper, we focus on the following six beliefs that are considered incorrect or undesirable in the literature about this measure:

(1) Alpha is equal to the reliability of a test score (e.g. Cho & Kim, 2015; Cortina, 1993; Sijtsma, 2009),
(2) The value of alpha is independent of the number of items of a test (e.g. Cortina, 1993),
(3) Alpha is an indication of the one-dimensionality of a test score (e.g. Cortina, 1993; Nunnally & Bernstein, 1994; Schmitt, 1996; Sijtsma, 2009),
(4) Alpha is the best choice among reliability coefficients (e.g. Cho & Kim, 2015; Sijtsma, 2009),

(5) There is a particular level of alpha that is desired or adequate (e.g. Cortina, 1993; Schmitt, 1996; Cho & Kim, 2015), and
(6) If removing an item increases alpha, the test is better without that item (Cho & Kim, 2015).

Next, we discuss these beliefs one by one.

The first undesirable belief is that alpha is equal to the classical *reliability* of a test score, which is, notably, not the same as the *internal consistency*. Alpha is only a proper (unbiased) estimate of reliability if the test items satisfy the model of essential tau-equivalence. Items are essentially tau-equivalent if each item measures the same latent variable, on the same scale, but with possibly different means and different errors (Graham, 2006; see also, Cho, 2016).). Another requirement is that the item errors in the model are uncorrelated (Green & Hershberger, 2000). However, essential tau-equivalence can usually not be assumed in which case alpha tends to underestimate the reliability of a test score. In contrast, if the assumption of uncorrelated item errors is violated, alpha may also overestimate reliability (Cho & Kim, 2015; Lucke, 2005).

The second undesirable belief is that the length of the test does not affect the value of alpha. If the average item covariance and variance are kept constant, however, alpha increases with the number of items (for a formal proof, see Appendix 2). This property makes sense since test scores usually have higher reliability if they are based on more parts. It may be that this property is usually not taken into consideration when alpha is reported and interpreted. It is also claimed that it is often not realized that tests that consist of a substantial number of items have a relatively large alpha simply because of the number of items (Cortina, 1993; Keszei, Novak, & Streiner, 2010; Tavakol & Dennick, 2011).

A third undesirable belief is that alpha is an indication of the one-dimensionality of a test score. A high value of alpha, however, does not guarantee that a test is measuring a single concept. Data can have two, three or more underlying factors, but with a sufficient number of items, the value of alpha can still be high in such cases (Schmitt, 1996; Sijtsma, 2009). Alpha is also not a measure of the extent to which there is a general factor present in the set of items and, therefore, the extent to which the items are interrelated (Cortina, 1993). It could be argued that this belief also confuses validity (does a test measure what it is intended to measure) and reliability.

The fourth arguably undesirable belief is that alpha is the best choice among reliability coefficients. Unlike the previous claims, there is no correct belief here: Of course, one could be of the opinion that alpha is indeed the most appropriate measure in some situations. Nevertheless, we think it is important to study researchers' awareness of alternatives. Various coefficients have been proposed that are valid measures of reliability under more general conditions than alpha (e.g. Omega (McDonald, 1999), which can be seen as a family of reliability coefficients, of which the unidimensional version, which Cho and Chun (2018) labeled congeneric reliability, seems to be most common. Another alternative is the greatest lower bound (glb; Woodhouse & Jackson, 1977). Since they can be used under more general conditions, these coefficients are considered by some to be better reliability alternatives than alpha (Cho & Kim, 2015; Revelle & Zinbarg, 2009; Sijtsma, 2009). These authors do not seem to agree on the usability of all these measures (e.g. Revelle & Zinbarg), however, amongst others because we do not yet fully understand how these estimates behave under specific conditions.

The fifth undesirable belief discussed here is that there is a particular level of alpha that is desired or adequate, independent of the context. The 'magical' cut-off value that separates good from bad tests is usually considered to be 0.70 or 0.80. It is often suggested that Nunnally (1978) proposed the 0.70 value, but in this work, he does not explicitly make this claim. More recently, researchers tend to see 0.80 as a minimum alpha (Cho & Kim, 2015). However, these values are not based on empirical research or logical reasoning per se, but seem more of a rule of thumb. There seems consensus among statisticians that a single criterion indicating a high reliability is unwanted, and, if a criterion is necessary in the first place, should be context dependent.

A sixth undesirable belief has to do with the justification for removing an item. Researchers are allegedly frequently motivated to remove items if this increases alpha (Cho & Kim, 2015; Cortina, 1993). Having obtained the desired level of alpha, researchers tend to use a test without further consideration of the dimensionality or criterion validity of the test. Furthermore, in the case of a low alpha value, items are often deleted using the 'alpha if item deleted'-statistic. However, using such an indiscriminant procedure may harm validity.

Misunderstandings of alpha can have serious implications for scientists personally (when, for example, their papers are rejected based on an incorrect understanding of alpha), but also for science as a whole. For example, people could be inclined to compromise the validity of their instruments by removing items, based on the incorrect understanding that alpha should have a certain minimal value. Similarly, others could try to add rephrased versions of current items of their questionnaire to boost alpha, or remove items that may lower alpha, despite the fact that this does not affect the quality of the questionnaire positively. Even though we are not claiming that researchers are regularly manipulating alpha to get nicer looking outcomes, we do believe that it is important that if researchers use alpha, they have a proper understanding of its meaning. We think the debate on alpha's use is too important to rely on alleging the prevalence of misunderstanding among researchers, and should be informed by actual empirical data instead.

Method

Procedure

Our sample consists of the corresponding authors of articles in which coefficient alpha was reported, published in 2011–2015 in nine top-tier journals in four different research disciplines, being psychology, management studies, pedagogy and public administration. We chose to focus on top-tier journals as we expected that researchers who publish in these journals are as least as well informed about coefficient alpha as their colleagues who do not use alpha, or who publish in lower ranked journals. We selected these four disciplines as interpretational problems surrounding coefficient alpha is frequently addressed in psychology, receives some attention in management studies and pedagogy, and is hardly discussed in public administration. In particular, we included the following 9 journals: *Journal of Management* (33 issues) and *Organizational Behavior and Human Decision Processes* (30 issues) in management studies; *Learning and Instruction* (30 issues) and *Child Development* (30 issues) in pedagogy; *Journal of Psychosomatic Research* (58 issues) and *Personality and Individual Differences* (80 issues) in psychology; and *The American Review of Public Administration* (31 issues), *Journal of Public Administration Research and Theory* (23 issues) and *Public Administration Review* (33 issues) in public administration. We only included corresponding authors in our sample as they (by definition) have indicated they are willing to answer questions concerning their research.

Using an R script, we downloaded all articles published in 2011–2015 from the selected journals. Next, we used the grep function in R to search in every downloaded publication for the application of coefficient alpha using the following case-insensitive search-string: 'Cronbach' OR 'coefficient alpha' OR 'internal consistency' OR 'internal consistencies'. We decided to exclude the psychometrical symbol for coefficient alpha (i.e. α) from the search string as the same symbol is also used to denote the required level of significance in statistical testing and – based on our own experiences – rarely used without one of the selected key terms when referring to coefficient alpha. In case at least one of the selected key terms was used in a publication, we extracted the name and email address of the corresponding author using the grep- and sub-functions in R.

We subsequently preprocessed our initial list of 2303 corresponding authors. First, names and email addresses that were lost in the mining process – due to slight deviations in the layout of certain issues – were manually retrieved from the selected articles. Second, we corrected all names whose spelling got corrupted in the extraction process, mostly caused by the presence of diacritical

marks. Third, all duplicate authors were removed from the initial list. We also excluded the authors of the current study. Our final sampling frame consisted of 1944 unique potential participants that had reported coefficient alpha in *Journal of Management* (103 authors), *Organizational Behavior and Human Decision Processes* (96 authors), *Learning and Instruction* (145 authors), *Child Development* (117 authors), *Journal of Psychosomatic Research* (247 authors), *Personality and Individual Differences* (1112 authors), *The American Review of Public Administration* (44 authors), *Journal of Public Administration Research and Theory* (39 authors) and *Public Administration Review* (41 authors).

Potential participants received a personalized email message sent in September 2016 containing a link to our online questionnaire programmed in Qualtrics. They were instructed to complete the questionnaire using their current knowledge (i.e. without use of any other sources of information). Potential participants were also informed that all data would be processed anonymously, even though they were given the opportunity to provide us with their email address in case they would like to be informed of the outcomes of the study afterwards.

Sample

Of the 1944 emails sent out to possible participants, 204 emails bounced because the email addresses were no longer active, meaning that we sent e-mails to 1740 researchers. Seven days after sending the initial invitation, a reminder was sent. After two weeks, participation for the survey was closed. We removed participants who completed less than a third of the questions, resulting in the removal of 49 participants. Eventually, a total of 664 participants answered at least a third of the questions, resulting in a response rate of 38%. Most researchers (n = 587, 88%) provided information about their field of research, with a few mentioning two affiliations. Out of those who filled in a field of research, Social sciences were most frequently indicated (n = 534, 91%). The rest of the researchers affiliated themselves with Medical sciences (n = 35, 6%), Statistics and methodology (n = 11, 2%) or Economical sciences (n = 10, 2%), and one person that we could not categorize in one of the above categories.

Questionnaire

The questionnaire consisted of 11 items. Eight of those were used to investigate the degree to which the six undesirable beliefs (as describes in the introduction of this paper) exist among scholars. A pilot study among colleagues confirmed that the questions of our survey were easy enough to understand. In order to minimize the risk of forced answers, we allowed respondents to skip questions. We, however, did not explicitly point out that possibility, as we feared it might lead to a decreased response rate. In case multiple answers were allowed, this was explicitly mentioned. The questionnaire has been added as Appendix 1.

Analysis

The data were analyzed by providing descriptive statistics per misinterpretation. The reason for not providing inferential statistics was because of the exploratory nature of the study. Given that questions could be skipped, percentages are given for those participants who answered the particular question. In addition, the association between the existence of the misinterpretations and the educational experiences of the respondents and their perceived level of understanding of alpha will be presented. The data and the scripts to analyze the data are publically available here: https://osf.io/32vfk/. To prevent others from being able to deduce data to individual participants, we excluded participants' field of research from the data file. Note that this variable was not combined with any other variable in our analyses.

Approval

This study was approved by the ethics committee of the Nieuwenhuis institute for Pedagogical and Educational Sciences, University of Groningen, the Netherlands.

Results

The frequencies of the six beliefs are presented below in separate tables. With respect to the first belief, it seems (see Table 1) that less than a third of the participants were aware of the fact that alpha is generally an *underestimate* of the reliability of the test at hand. The percentages do not seem to deviate that far from what one would expect if the answers were a result of random guessing.

Researchers seemed more aware of the fact that when the number of items is increased (in this case by artificially doubling the items), alpha increases too, with 82% of the respondents selecting this answer (see Table 2).

However, the fact that alpha is not an indication of the one-dimensionality (belief 3) seems to be less well known: only a fifth of the participants (see Table 3) correctly indicated that adding two groups of relatively highly correlated questions results in a rather high value of alpha, even when between-group correlations are zero.

Fourth, we were interested in whether researchers consider alpha the best choice among reliability coefficients. Since knowing an alternative in the first place is a prerequisite for answering this question, we first asked participants for their awareness of alternatives. A majority of 77% (n = 500 out of 653) indicated knowing an alternative and was able to mention at least one alternative to alpha: test-retest (164 times, 33%), split-half (132 times, 26%) and omega (74 times, 15%) were most frequently mentioned. Sixty percent (out of 500) indicated that they published at least one of these alternative measures in a scientific paper.

The fifth belief we were interested in had to do with researchers' use of a specific criterion for a sufficient or desirable level of alpha. A large majority (76%; see Table 4) indeed indicated using such a criterion. However, on closer inspection, out of the 501 participants who answered 'Yes,

Table 1. The answers (n = 590) to the question for undesirable belief 1 (coefficient alpha is equal to the reliability of a test). The percentage for the answer that is considered correct is presented bold.

Question	Answer	Percentage
Suppose we find an alpha of 0.78 in a large random sample of subjects who filled in a questionnaire. What does that mean about the reliability of the questionnaire?	Probably lower than 0.78	39%
	Probably is exactly 0.78	33%
	Probably higher than 0.78	**28%**

Table 2. The answers (n = 622) to the question for undesirable belief 2 (The alpha coefficient is independent of test length). The percentage for the answer that is considered correct is presented bold.

Question	Answer	Percentage
A test consists of 10 items and has a given alpha, e.g. 0.60. Suppose we double the test length afterwards by copying and pasting the responses to the test, thus making it a 20-item test, with every item occurring twice. What do you think happens to the value of alpha?	Alpha decreases	4%
	Alpha remains the same	14%
	Alpha increases	**82%**

Table 3. The answers (n = 534) to the question for undesirable belief 3 (The alpha coefficient provides the user with an indication of the degree of one-dimensionality of a test). The percentage for the answer that is considered correct is presented bold.

Question	Answer	Percentage
Suppose there are two groups of 10 items each. The correlations between items within each group are 0.60, but correlations between items from different groups are 0.00. [...] Select the value that you consider most correct.	0.10	36%
	0.50	44%
	0.90	**20%**

Table 4. Undesirable belief 5: a particular level of alpha is sufficient/desirable (n = 657).

Question	Answer	Percentage
Suppose you read aresearch article in which alpha is reported. Is there a critical value of alpha that you consider a generally applicable indication of a high reliability	Yes, namely …	76%
	No	24%

Table 5. The answers to the question under which conditions removing an item would be justifiable (undesirable belief 6; n = 576).

Reasons for removing an item *(multiple answers possible)*	Percentage
If this increases alpha	20%
If alpha increases with at least … *(value).*	16%
Provided that removing the item is reported in the publication	62%
Provided that the test length remains sufficiently long	30%
Other	38%

namely…', 481 provided at least a value or an explanation. Out of those, 78 (16%) indicated that they thought such a criterion should be context dependent, and another 26 (5%) provided a range rather than a single criterion. Out of the 381 (79%) researchers claiming to use a single criterion, 0.70 was mentioned 182 times (48%) and 0.80 was mentioned 118 (31%) times.

The sixth belief we discuss is the extent to which people consider it justified to remove an item based on the outcome of alpha for all items. It was found (see Table 5) that merely removing an item to increase alpha was not too often endorsed (20% without specifying the increase and 16% when alpha was to increase with at least a certain amount). About a third (30%) considers it okay to remove items provided that the test remains sufficiently long. For those who checked the 'Other' box (n = 219), 46% indicated that for removing content-related arguments are needed as well.

Lastly, the participants were also presented questions about their general use of alpha or its alternatives, and their experience with the technique. A large majority (82%; 542 out of 657) indicated that when confronted with an alternative reliability measure unknown to them in a paper they were to review, they would search for information about this measure. None of the participants would suggest to replace this measure with alpha, although 9% indicated they would suggest to add alpha.

On average, the participants scored somewhere in the middle on a 10-point scale indicating whether alpha was explained in their education (mean 5.4, standard deviation 2.6), and whether they considered themselves experts with respect to alpha (mean 5.9, standard deviation 1.7). The Pearson correlation between both measures and the number of incorrect answers for the first three beliefs, for which there clearly is an incorrect answer, was rather low (0.16 and 0.14, respectively). Based on these results, self-reported expertise and experience with alpha do not seem to prevent it from being misinterpreted.

In Table 6, researchers' reasons for presenting alpha in at least one of their papers (note that the participants were selected for having at least one publication in which alpha had been used) are presented. The most frequently mentioned reasons seem to indicate conformity, rather than a deliberate choice for the technique.

Table 6. Researchers' reasons for presenting alpha in at least one of their papers (n = 664).

Reasons for reporting alpha *(multiple answers possible)*	Percentage
Because I thought this would be required by the journal/the reviewer	53%
Because I was taught that this is what you should do when reporting a questionnaire	43%
Because I thought that this is what I should do when reporting a questionnaire	40%
Because it is common practice in my area of expertise	74%
Because I was unfamiliar with alternative measures	12%
Other	14%

Conclusion and discussion

Alpha is claimed to be misunderstood frequently (e.g. Cho & Kim, 2015; Green, Lissitz & Mulaik, 1977; Sijtsma, 2009; Schmitt, 1996). As far as we know, however, our paper is the first to present actual data of researchers' beliefs surrounding alpha. Based on its results, we are able to confirm some of the claims, but also provide a more balanced picture for others.

It was found that only a small minority (28%) of researchers in our sample seemed aware of the fact that alpha is to be considered an *under*estimate of reliability. Although this may sound low, some nuancing is necessary: In some cases, alpha may in fact be an overestimate (Cho & Kim, 2015; Lucke, 2005; Raykov & Marcoulides, 2017), if the assumptions that the error terms do not correlate does not hold. Although we still think that the best defensible option is that alpha is *probably* an underestimate, this depends on what one considers the likelihood of a strong violation of uncorrelated item errors. In case one thinks this likelihood is rather high, stating that alpha is probably an *under*estimate may in fact be a reasonable and justified answer. From the current answers, it cannot be determined what considerations caused the participants to answer in a certain way. For that reason, considering the overestimate answer incorrect might be premature, but nevertheless we think it is interesting to see the variability of answers for this particular question. In fact, quite a large group (39%) expects alpha to be an *over*estimate. Moreover, we found that most believed that alpha would not be high for a clearly two-dimensional scale, although it can be proven (see Appendix 3) that this is in fact the case.

We did not find that many incorrectly believed that alpha is unaffected if you make a test longer with similar items. In fact, a clear majority (82%) seemed aware that doubling a questionnaire with the same questions (thus not adding any information to the questionnaire) does increase the value of alpha. Moreover, we found that most researchers (77%) were aware of alternative measures for internal consistency, and a large proportion (60%) had indeed reported one of those in at least one published paper. Lastly, most researchers claimed to be hesitant to remove items to increase alpha, and none would advise someone to replace an alternative measure by alpha when reviewing a paper.

All in all, we found support for claims that alpha is misunderstood regularly, as many (Cho & Kim, 2015; Cortina, 1993; Sijtsma, 2009) had anticipated. Nevertheless, the picture might be a bit more positive than some expected, assuming that our findings are representative of the actual misunderstanding in practice.

A few remarks are in place. First of all, our sample cannot be considered a random sample of all researchers. Since we selected people who already reported alpha in a publication, our sample might arguably be more knowledgeable than the average researcher. That is, misunderstandings among all researchers could be more widespread than our data suggest. On the other hand, we think it is more relevant to get a grasp of the understanding of researchers who actually use the measure, than of those who do not. In that sense, the fact that our sample is not representative for the academic community as a whole might not be problematic.

Second, participants could have tried to give the answers they expected we wanted to hear, rather than the answers that most reflected their beliefs. However, the displayed amount of misunderstanding suggests otherwise. Nonetheless, especially for more open questions, social desirability might have affected the outcomes somewhat.

Third, due to the exploratory nature of this study, we intentionally kept the questionnaire short. Naturally, this affected the amount of questions we could ask, and the depth of our questions, On the other hand, because of the limited length, we had a fairly decent response rate given this type of research, and we managed to provide empirical evidence for claims that were unsubstantiated so far. So, although we acknowledge that asking more specific questions (rather than the more general questions we asked) would have been valuable, we think that the exploratory nature of the study justifies the type of questions we asked. Naturally, we encourage others to attribute to these first steps on this relatively uncharted territory.

To summarize, our data clearly suggest that alpha is often misunderstood, although there seem be some indications that there is at least some awareness of its pitfalls. Our results also seem to suggest (but this needs more extensive studying) that the reason for its current popularity is a result of its popularity in the past, rather than of its users' enthusiasm about or even awareness of its properties. Those who would like to change the current popularity of coefficient alpha seem to have a rather difficult task, since the past is not easily changed. More knowledge of the prevalence, as we presented in this paper, seems a necessary and so far missing step for changing how researchers seem to deal with reliability.

Suggestions

So where to go from here? Should our paper be seen as supporting the claim that we should abandon coefficient alpha altogether, because it is indeed regularly misinterpreted? We think such a conclusion would be unwarranted: This should at least also be dependent on its usability. Unfortunately, we seem to be in a situation in which the debate about coefficient alpha's usefulness has not been settled yet. Instead of adding our opinion to this debate, we think proposing the following suggestions for both substantive researchers and experts is a more constructive approach. According to us, the following steps should be taken in order to find a way out of the current stalemate.

Suggestions for potential users

As long as experts on coefficient alpha do not seem to agree, making a decision on whether to use alpha or not may not be easy. If one decides to report alpha, reading up on its potential and its limitations, rather than copying how others are using it, seems pivotal. Although the latter approach may be attractive from a time management perspective, the frequency of misinterpretations as shown in this paper is a warning that betting on others' understanding may not be safe. Second, it may be good advice to add one or more alternative measures as well. Although reporting multiple measures may seem abundant, it at least makes the outcomes accessible for both proponents and opponents of coefficient alpha. Third, one should keep in mind that whatever measure is used for reliability, reliability is only a prerequisite for validity. So, independent of the measure at hand, a relatively high value should by no means be seen as an excuse not to discuss the validity of the test or questionnaire at hand.

Although we focused on coefficient alpha since it is the most frequently used measure to estimate reliability, we acknowledge that not all its features, some of which are regularly misunderstood as we have shown, are unique. For example, the increase of alpha when the number of items increases is also found in many other measures. Therefore, we want to stress that even if users would replace alpha by another coefficient, they should still be wary of potential misinterpretations. A thorough understanding of the measure at hand is necessary anyway.

Suggestions for psychometricians and other reliability experts

The debate on coefficient alpha seems rather polarized at the moment, and, unfortunately, seems mostly ignored by substantive researchers. It would be beneficial if this discussion would either be settled, or if there would be more clarity in which situations and under which assumptions alpha or its alternatives seem preferable. For this, more research on the properties of alternative measures, resulting in clear guidelines for researchers, is probably essential. Second, if we are to adopt alternatives for, or additions to, coefficient alpha, it would be helpful if there were agreements amongst alpha's opponents. Although not agreeing on an alternative does not in itself disqualify the arguments against alpha, it does make it understandable that until agreement is reached, alpha is still frequently used. Third, Kruyen and Van Assen (2018) suggested context-dependent rules of thumb. Adopting these may be helpful to help substantive researchers

evaluating their outcomes. The same may hold for Cho's (2016, p. 18) stepwise guide for choosing a reliability coefficient.

Suggestions for teachers and educators

As we have shown, alpha is regularly misinterpreted, as was predicted by many psychometricians. Given that alpha is still frequently used, and given that even if we were to stop using alpha starting today like McNeish (2017) seems to be advocating, the scientific literature is still filled with values of coefficients alpha, it is crucial that students are made aware of correct interpretations, and also of potential pitfalls. We think that this paper might be a helpful source when teaching about reliability coefficients. First of all, it provides information about what alpha is and what it isn't, and secondly the questionnaire as presented in Appendix 1 can be useful to test students' understanding.

These are just a few first suggestions to move us forward in this debate that has been going on for decades. Given the duration of the debate, and given the frequency of misunderstanding as presented in this paper (although it may not be as dramatic as some suspected), we are in dire need of progress regarding our use of measures of reliability.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

R. Hoekstra is an assistant professor at the University of Groningen whose research mainly focuses on how researchers understand and use statistical outcomes, and how the scientific context influences why they use it in a certain way. He has published regularly about use and understanding of significance testing and its alternatives.

J. Vugteveen is a PhD candidate at the University of Groningen. In her PhD project, she focusses on the validity of questionnaires aimed at measuring behavioral problems and how to improve implementation of these questionnaires in practice.

M. J. Warrens is an associate professor at the University of Groningen. His research interests are the development and comparison of indices for assessing similarity between individuals or objects. These indices are used in many different scientific disciplines, e.g. Epidemiology, Psychology, Educational Research, Ecological biology, Test theory, Market research and Remote sensing. His research produces guidelines on which indices to use and how to interpret them.

P. M. Kruyen is an assistant professor at the Radboud University in Nijmegen. His research focuses on civil servants' behavior, psychological characteristics and competences, and on how their work is affected by managerial techniques and organizational structures

ORCID

R. Hoekstra http://orcid.org/0000-0002-1588-7527
J. Vugteveen http://orcid.org/0000-0002-8098-4120
M. J. Warrens http://orcid.org/0000-0002-7302-640X
P. M. Kruyen http://orcid.org/0000-0003-0109-1744

References

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods*, *19*(4), 651–682.

Cho, E., & Chun, S. (2018). Fixing a broken clock: A historical review of the originators of reliability coefficients including Cronbach's alpha. *Survey Research*, *19*(2), 23–54.

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*, 207–230.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.
Field, A. (2009). *Discovering statistics using SPSS*. (3nd ed.). Los Angeles: Sage.
Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics. An introduction*. Los Angeles: Sage.
Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. What they are and how to use them. *Educational and Psychological Measurement*, *66*(6), 930–944.
Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling*, *7*, 251–270.
Green, S. B, Lissitz, R. W, & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality1. *Educational and Psychological Measurement*, *37*(4), 827-838.
Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255–282.
Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, *6*, 153-160.
Jackson, R. W. B., & Ferguson, G. A. (1941). *Studies on the reliability of tests*. Department of Educational Research, University of Toronto.
Keszei, A. P., Novak, M., & Streiner, D. L. (2010). Introduction to health measurement scales. *Journal of Psychosomatic Research*, *68*, 319–323.
Kruyen, P., & Van Assen, M. A. L. M. (2018). *On the Reliability of Surveys and Questionnaires in Public Administration Research: How Should We Care?* Unpublished manuscript.
Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160.
Lucke, J. F. (2005). "Rassling the hog": The influence of correlated item error on internal consistency, classical reliability and congeneric reliability. *Applied Psychological Measurement*, *29*, 106–125.
McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, New Jersey: Erlbaum.
McNeish, D. (2017). Thanks coefficient alpha, We'll take it from here. *Psychological Methods*. Advance online publication. doi: 10.1037/met0000144
Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw Hill.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill, Inc.
Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*, 343–355.
Raykov, T., & Marcoulides, G. A. (2017). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement*. doi:10.1177/0013164417725127
Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*, 145–154.
Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350.
Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120.
Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55.
Warrens, M. J. (2015). Some relationships between Cronbach's alpha and the Spearman-Brown formula. *Journal of Classification*, *32*, 127–137.
Warrens, M. J. (2016). A comparison of reliability coefficients for psychometric tests that consist of two parts. *Advances in Data Analysis and Classification*, *10*, 71–84.
Winer, B. L. (1971). *Statistical principles in experimental design* (2nd). New York: McGraw-Hill.
Woodhouse, B., & Jackson, P. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika*, *42*(4), 579–591.

Appendix 1. Questionnaire

Coefficient alpha (also known as Cronbach's alpha) is a measure that is regularly used within empirical research. The following short questionnaire is about researchers' intuitions and ideas about it. The questionnaire consists of 11 questions. We kindly ask you to answer these questions based on your current knowledge and ideas (no other sources of information needed). Please indicate for each of the items which one of the statements reflects your ideas about coefficient alpha by checking the box next to that particular statement. All data will be processed anonymously. The questionnaire will take no longer than 5 minutes.

Thank you in advance!
Sincerely,
Rink Hoekstra, Matthijs Warrens, Peter Kruyen, & Jorien Vugteveen

1. You've used alpha in at least one of your research papers. Why did you choose to report this particular measure of reliability? (multiple answers possible)

- □ Because I thought this would be required by the journal/the reviewer
- □ Because I was taught that this is what you should do when reporting a questionnaire
- □ Because I thought that this is what I should do when reporting a questionnaire
- □ Because it is common practice in my area of expertise
- □ Because I was unfamiliar with alternative measures
- □ Other: _____

2a. Do you know any other reliability measures of a test score besides alpha?

- ○ No
- ○ Yes, namely _____

If No Is Selected, Then Skip To Question 3
2b. Have you ever reported this other (or one of these other) reliability measure(s) in a scientific paper?

- ○ Yes
- ○ No

3. Suppose you are asked to review a paper in which a reliability measure other than alpha is presented, a measure that you are not familiar with. What would you do?

- ○ I would suggest to replace this measure with alpha.
- ○ I would suggest to add alpha.
- ○ I would search for information about the measure, in order to be able to have an informed opinion about its usability.
- ○ I would probably not say anything about the measure in my review.
- ○ Other: _____

4. Suppose you read a research article in which alpha is reported. Is there a critical value of alpha that you consider a generally applicable indication of a high reliability?

- ○ No
- ○ Yes, namely _____

5. Suppose we find an alpha of 0.78 in a large random sample of subjects who filled in a questionnaire. What does that mean about the reliability of the questionnaire?

- ○ The reliability is probably lower than 0.78.
- ○ The reliability is exactly 0.78.
- ○ The reliability is probably higher than 0.78.

6. A test consists of 10 items and has a given alpha, e.g. 0.60. Suppose we double the test length afterwards by copying and pasting the responses to the test, thus making it a 20-item test, with every item occurring twice. What do you think happens to the value of alpha?

- ○ Alpha decreases.
- ○ Alpha remains the same.
- ○ Alpha increases.

7. Suppose there are two groups of 10 items each. The correlations between items within each group are 0.60, but correlations between items from different groups are 0.00. Below we list several possible values of alpha. Select the value that you consider most correct.

- ○ 0.10
- ○ 0.50
- ○ 0.90

8. Suppose alpha increases if an item is deleted. Below we list several possible situations in which it may or may not be justifiable to remove the item. Select all situations in which you agree with removing an item.

- ☐ Remove the item if this increases alpha.
- ☐ Remove the item if alpha increases with at least … (*in case you select this item, insert value*). _____
- ☐ Remove the item provided that removing the item is reported in the publication.
- ☐ Remove the item provided that the test length remains sufficiently long.
- ☐ Other: _____

9. To what extent was alpha explained in your own education?

- ○ 0
- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ 7
- ○ 8
- ○ 9
- ○ 10

10 How experienced do you consider yourself with respect to the understanding of alpha?

- ○ 0
- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ 7
- ○ 8
- ○ 9
- ○ 10

11. What is your field of research?

- ○ Social sciences
- ○ Medical sciences
- ○ Economical sciences
- ○ Statistics and methodology
- ○ Other: _____

Thank you for your cooperation!
In case you're interested in the results of this study, please insert your e-mail address. Of course, your e-mail address will only be used for the purpose of sharing the results of this study with you. Your e-mail address will not be linked to your response to this questionnaire. All responses will be processed anonymously.

Appendix 2

The second question of the questionnaire was used to assess the incorrect belief that alpha is independent of the number of items. In the scenario presented in Question 2, alpha will always increase, regardless of the particular data at hand. A proof is as follows. Suppose there are k items. Denote the sum of the item variances by $v = \sum_{i=1}^{k} \sigma_i^2$ and twice the sum of the item covariances by $c = 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \sigma_{ij}$. Alpha is then given by $\alpha = kc/(k-1)(v+c)$. Furthermore, after doubling the items, the new alpha is given by $\alpha^* = k(v+2c)/(2k-1)(v+c)$. After some

algebra, we find that we have $\alpha < \alpha^*$ if and only if $c(2k-1) < (k-1)(v+2c)$ or equivalently $v + c < kv$. The latter inequality follows from the fact that the average covariance between the items never exceeds the average variance of the items (e.g. Winer, 1971).

Appendix 3

The third question of the questionnaire was used to assess the incorrect belief that alpha is an indication of the one-dimensionality of a test score. In the scenario presented in Question 3, alpha is likely to produce a value close to '0.90'. The actual value depends on the variances and covariances of the data at hand. An approximation value can be obtained using standardized alpha, which only requires the average correlation between the items. Suppose there are k items. Let $\bar{r}$ denote the average correlation. Standardized alpha is then given by $\alpha_s = k\bar{r}/(1 + (k-1)\bar{r})$. For $k = 10$ and $\bar{r} = 0.60$, we obtain $\alpha_s = 0.94$, which is the value of standardized alpha corresponding to either group of 10 items in the scenario of Question 3. The average correlation between the 20 items in this scenario is 0.284. For $k = 20$ and $\bar{r} = 0.284$, we obtain $\alpha_s = 0.89$, which is the value of standardized alpha corresponding to the 20 items in the scenario of Question 3.